УДК 681.322.01/.08 + 681.3.058/.069

А.Б. Бессонов
(Уральская государственная
лесотехническая академия)

УСОВЕРШЕНСТВОВАННЫЙ МЕТОД ИНВЕРТИРОВАННОГО ФАЙЛА

В большинстве существующих систем управления базами данных используются такие модели данных, в которых связи между ними хранятся вместе с самими данными. Например, указатели связи между записями входят в состав записей. Объединение элементов данных в записи реализуется путем их физически последовательного размещения.

Однако, при всевозрастающей сложности баз данных, такой подход не всегда является эффективным для организации быстрого поиска данных. Существует ряд систем управления базами данных (СУБД), поддерживающих раздельное хранение данных и их связей.

Основными причинами раздельного хранения данных и связей являются обеспечение более полной независимости данных и возможности ускорения поиска.

В практической реализации раздельного хранения данных и связей часто используется метод инвертированного файла, позволяющий осуществлять быстрый поиск для запросов общего вида, включающих спецификацию значений вторичного индекса.

В некоторых СУБД, таких как ADABAS, предусмотрены средства организации полностью инвертированного файла. Опыт создания БД в ADABAS показал, что метод инвертированного файла позволяет эффективно обрабатывать только простые запросы без сложных булевых функций. Необходимость

совершенствования метода инвертированного файла очевидна.

Для решения задачи модификации метода инвертированного файла была создана модель организации данных в СУБД ADABAS.

Рассмотрены три множества T*, D*, H*. При этом, что множество D* есть гомоморфный образ множества T*. При отображении F множество D* есть гомоморфный образ множества H* при отображении S. Задаются отношение эквивалентности V, определяемое в множестве T* отображением F : F(h)=F(q) (h, q ∈ T*), и отношение эквивалентности Q, определяемое в множестве H* отображением S : S(x)=S(y) (x, y ∈ H*). Классы эквивалентности в T* и H* образуются множеством элементов, имеющих один и тот же образ в D*. Отношение V совместимо с любым внутренним бинарным законом композиции, заданным на T*. Действительно, пусть h1, h2,, q1, q2 ∈ T* такие, что h1≡h2(mod V), q1≡q2(mod V) или F(h1)=F(h2)=d1, F(q1)=F(q2)=d2, тогда h1+q1=h2+q2(mod V), так как по определению гомоморфизма F(h1+q1)=F(h1)+F(q1)= =d1xd2=F(h2)xF(q2)=F(h2xq2), где + и x внутренние бинарные законы композиции, заданные на T* и D* соответственно.

Соответствие между отображениями S и F определяется следующим выражением: ∀(h∈T)∃(d∈D*)[S-1F(h)={x|x∈H*}]Λ Λ∀(x∈H*)∃(d∈D*)[F-1S(x)={h|h∈T*}].

Определены операции, образующую алгебру, порожденную состоянием базы данных.

На каждом множестве T*, D*, H* определен внутренний бинарный закон композиции – умножение, эквивалентность цепочек.

Введено понятие формальной системы L на базовом множестве H* как любое подмножество свободной полугруппы G, образующими которой являются элементы множества H*.

Определена операция умножения L1 и L2 как операция, ставящая им в соответствие $L=\{\alpha\varphi|\alpha\in L1, \varphi\in L2\}$

Подстановку символов вместо символов h1, … hn системы L, систем L1, … Ln определим как операцию, сопоставляющую L на базовом множестве $H=\{h1, \ldots, hn\}$ и L1, … Ln на базовых множествах H1, …, Hn соответственно, следующую систему на базовом множестве $H1\cup \ldots \cup Hn$

$L'\cup\{\alpha i1 \ldots \alpha ik \mid hi1 \ldots hik \in L, \alpha i1 \in Li1, \ldots, \alpha ik \in Lik\}$

Вводится понятие отношения как множество $V\subseteq T^*$, если выполняется следующее выражение: $\forall(\alpha\in V)\ [F(\alpha)=\gamma, \gamma\in L(I`/I)]$. Множество V можно получить, зная T*, D* и F.

Цепочку $\beta\in H^*$ предлагается называть схемой отношения V, если $\forall(\alpha\in V)\ [F(\alpha)=\gamma \wedge S\text{-}1(\gamma)=\beta, \gamma\in L(I`/I), \beta\in H^*]$.

Множество H'={V|V – отношение, $V\in T^*$} будем называть базой данных, а множество схем, определяющих базу данных, – схемой базы данных.

Введенные выше определения позволяют рассматривать базу данных как формально-логический объект и позволяют сформулировать общее правило табличного представления результатов умножения конечного числа доменов: произведение элементов $\alpha1$ и $\alpha2$, стоящих на местах (i,j) и (k,i), должно находиться на месте (k,j).

Существенным свойством такой таблицы является то, что номер строки первого операнда однозначно определяет номер столбца, где расположен второй операнд операции умножения, а это в случае поиска информации приводит к сокращению времени поиска записей.

Основным недостатком таблицы умножения является ее большой размер. К примеру, если заданы 2 домена по 2 элемента и схема отношений, то все возможные записи отношения представляются в таблице размерностью 3x3.

Однако таблицу умножения не обязательно хранить в области описания свойств данных, достаточно зафиксировать и хранить ее обобщенные свойства. Для этого предлагаются следующие алгоритмы формирования указателей W1, W2, W3, W4, обеспечивающие представление этих свойств.

Алгоритм W1. Выполняется просмотр строк первого столбца таблицы умножения. Если в j-й строке расположено αi, в j-й строке второго столбца указателя W1 располагаются номера первой и последней записей в упорядоченном отношении V1, где αi является первым операндом (или левой частью).

Алгоритм W2. Выполняется просмотр элементов доменов отношения V в соответствии со схемой отношения V.

Для i-го домена заполняется i-я строка указателя W2. Для каждого значения hj, принадлежащего i-му домену, выполняется просмотр первого столбца таблицы умножения. На (i,j)-м месте указателя W2 помещается номер строки первого столбца таблицы умножения, в которой первый раз встречается второй операнд умножения (или первая часть), оканчивающийся j-м элементом i-го домена. Последний элемент каждой строки W2 равен номеру строки первого столбца таблицы умножения, следующей за последним операндом, оканчивающимся последним значением i-го домена.

Алгоритм 3. Просматриваются столбцы, начиная со второго, таблицы умножения отношения V. Для каждого столбца формируется список пар чисел, образующих таблицу W31. Каждая пара указывает номер первой и последней строки, оканчивающихся одинаковыми значениями. Разбиение списка пар чисел, представленных в виде таблицы W31, осуществляется элементами таблицы W32. i-я строка W32 характеризует i-й столбец таблицы умножения и указывает начало списка для i-го столбца, а i+1-й элемент будет указывать окончание этого списка.

Алгоритм W4. Выполняется просмотр i-го столбца таблицы умножения отношения V и вычисляются W4(i,1), как номер строки первого столбца таблицы умножения, где первый раз встретилось значение i-го домена. W4(i,2) есть количество значений всех доменов за i-м, согласно схеме отношения V, т.е.

$$W_4(i,2) = \sum_{j=j+1}^{n} m_j$$

где mj – количество элементов в j-м домене;

W4(i,3)=W4(i-1,3)+W4(i-1,2)*(W4(i,1)-W4(i-1,1)) для i>1;

W4(n,3)=0 для i=n;

W4(i,4)=0, для i=1;

W4(i,4)=W4(i-1,4)+mi-1.

Предлагаемые указатели являются улучшением метода инвертированного файла, поскольку предлагаемый инвертированный файл состоит не только из многоуровнего индекса, но и из набора указателей, обеспечивающих доступ к записям данных в соответствии с определенным критерием ключевого поля. Сравнительная характеристика результатов исследования приведена ниже.

Характеристики результатов исследования (в БД 1000 записей)

№ п/п	Показатель	Значение
1	Время поиска и выборки записи при стандартной для ADABAS организации инвертированного файла, мс	4
2	Время поиска и выборки записи по предлагаемому методу организации инвертированного файла, мс	3,24